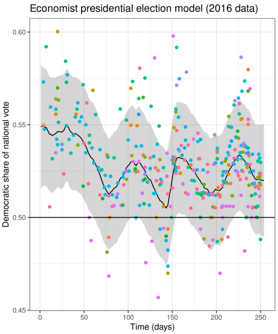
Approximate data deletion and replication with the Bayesian influence function

Ryan Giordano (rgiordano@berkeley.edu, UC Berkeley), Tamara Broderick (MIT)

MIT Robustness and Influence Functions Workshop

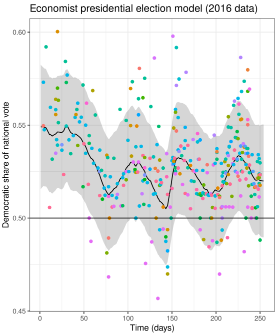
Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \ldots, x_N =$ polling data.
- $\theta =$ parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with $p(X|\theta)$
- $f(\theta) =$ Democratic % of vote on election day

Economist 2016 Election Model [Gelman and Heidemanns, 2020]

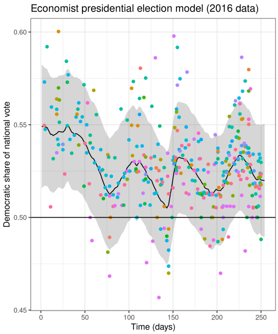


A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \ldots, x_N =$ polling data.
- $\theta =$ parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with $p(X|\theta)$
- $f(\theta) =$ Democratic % of vote on election day

Many values of θ are consistent with the data.
How can we account for our uncertainty when making election predictions?

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



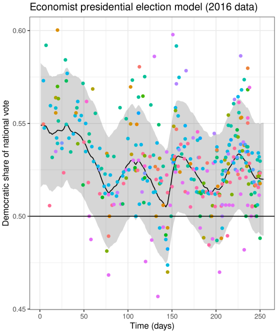
A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \ldots, x_N =$ polling data.
- $\theta =$ parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with $p(X|\theta)$
- $f(\theta) =$ Democratic % of vote on election day

Many values of θ are consistent with the data.
How can we account for our uncertainty when making election predictions?

Bayesian answer: Average over values that are plausibly consistent with the data.

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \ldots, x_N$ = polling data.
- θ = parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with $p(X|\theta)$
- $f(\theta)$ = Democratic % of vote on election day

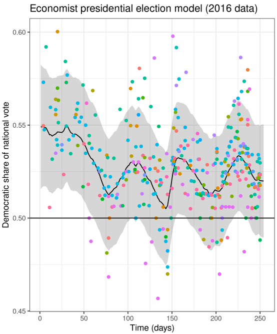
Many values of θ are consistent with the data.
How can we account for our uncertainty when making election predictions?

Bayesian answer: Average over values that are plausibly consistent with the data.

Formally: Define a prior $p(\theta)$, compute the posterior $p(\theta|X)$, and estimate

$$f(\theta) \approx \mathop{\mathbb{E}}_{p(\theta|X)} [f(\theta)], \text{ with uncertainty proportional to } \sqrt{\mathop{\mathrm{Var}}_{p(\theta|X)} (f(\theta))}.$$

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \ldots, x_N$ = polling data.
- θ = parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with $p(X|\theta)$
- $f(\theta)$ = Democratic % of vote on election day

Many values of θ are consistent with the data.
How can we account for our uncertainty when making election predictions?

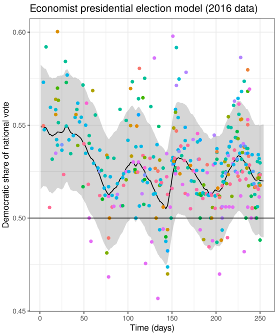
Bayesian answer: Average over values that are plausibly consistent with the data.

Formally: Define a prior $p(\theta)$, compute the posterior $p(\theta|X)$, and estimate

$$f(\theta) \approx \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)], \text{ with uncertainty proportional to } \sqrt{\mathop{\mathrm{Var}}_{p(\theta|X)}(f(\theta))}.$$

Practically: We compute Markov chain Monte Carlo (MCMC) draws from $p(\theta|X)$.

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



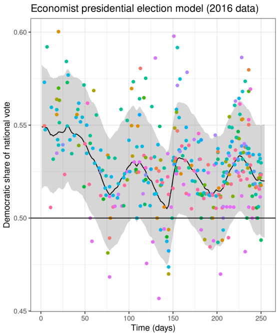
A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

We want to know $\underset{p(\theta|X)}{\mathbb{E}} [f(\theta)]$.

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

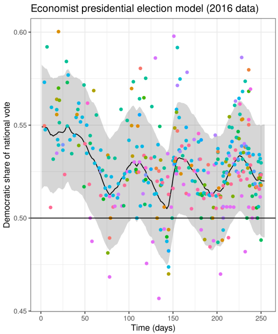
- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

The people who responded to the polls were randomly selected.
If we had selected a different random sample, how much would our estimate have changed?
How can we estimate $\underset{X \overset{iid}{\sim} \mathbb{F}}{\mathrm{Var}} \left(\underset{p(\theta|X)}{\mathbb{E}} [f(\theta)] \right)$?

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

The people who responded to the polls were randomly selected.
If we had selected a different random sample, how much would our estimate have changed?
How can we estimate $\underset{X \overset{iid}{\sim} \mathbb{F}}{\mathrm{Var}} \left(\underset{p(\theta|X)}{\mathbb{E}} [f(\theta)] \right)$?

- Except in special cases, we expect $\underset{p(\theta|X)}{\mathrm{Var}} (f(\theta)) \neq \underset{X \overset{iid}{\sim} \mathbb{F}}{\mathrm{Var}} \left(\underset{p(\theta|X)}{\mathbb{E}} [f(\theta)] \right)$
- We are interested re–sampling for *this* election, not a hypothetical future election

Results

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Problem: Each MCMC run takes about 10 hours (Stan, six cores).

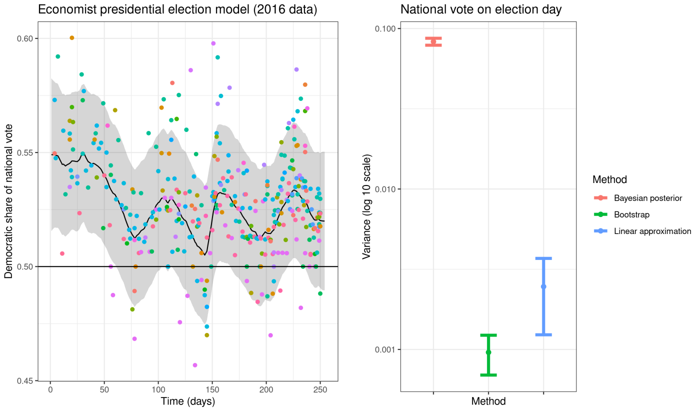
Proposal: Use full–data posterior draws to form a linear approximation to *data reweightings*.

Results

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Problem: Each MCMC run takes about 10 hours (Stan, six cores).

Proposal: Use full–data posterior draws to form a linear approximation to *data reweightings*.

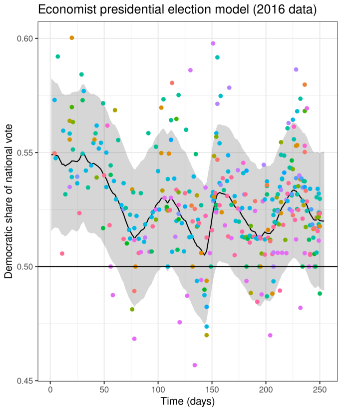


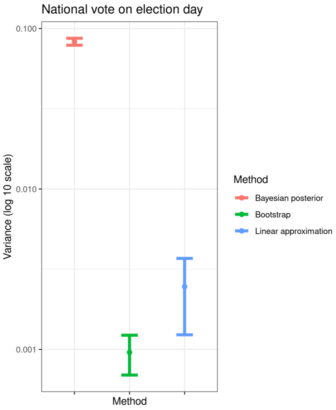
Results

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Problem: Each MCMC run takes about 10 hours (Stan, six cores).

Proposal: Use full–data posterior draws to form a linear approximation to *data reweightings*.

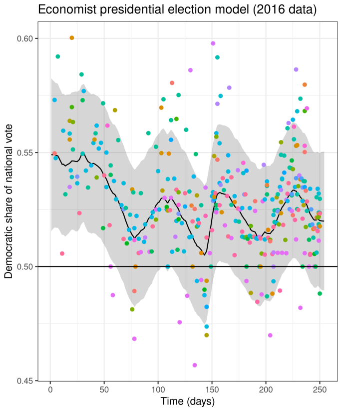


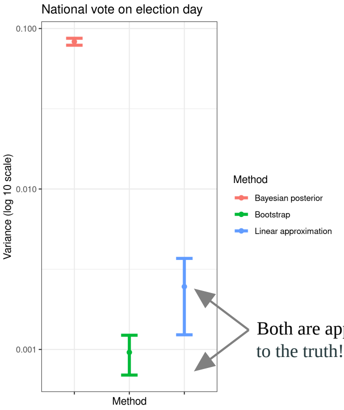


Compute time for 100 bootstraps:	51 days
Compute time for the linear approximation: (But note the approximation has some error)	Seconds

Results

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]
Problem: Each MCMC run takes about 10 hours (Stan, six cores).
Proposal: Use full–data posterior draws to form a linear approximation to *data reweightings*.





Both are approximations to the truth!

Compute time for 100 bootstraps: 51 days

Compute time for the linear approximation: Seconds
(But note the approximation has some error)

Data re-weighting.

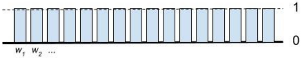
Augment the problem with *data weights* $w_1, \dots, w_N$.

Data re-weighting.

Augment the problem with *data weights* $w_1, \ldots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \qquad \Rightarrow \text{We write } \mathop{\mathbb{E}}_{p(\theta|X,w)} [f(\theta)] .$$

Original weights:

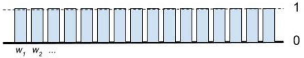


Data re-weighting.

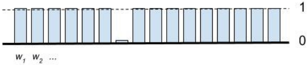
Augment the problem with *data weights* $w_1, \ldots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \quad \Rightarrow \text{We write} \underset{p(\theta|X,w)}{\mathbb{E}} [f(\theta)] .$$

Original weights:



Leave-one-out weights:

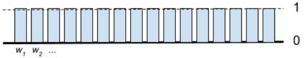


Data re-weighting.

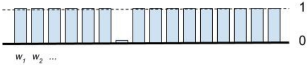
Augment the problem with *data weights* $w_1, \ldots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \quad \Rightarrow \text{We write} \underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)].$$

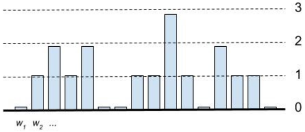
Original weights:



Leave-one-out weights:



Bootstrap weights:

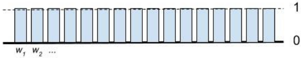


Data re-weighting.

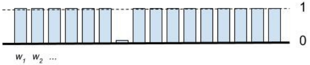
Augment the problem with *data weights* $w_1, \ldots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \quad \Rightarrow \text{We write} \underset{p(\theta|X,w)}{\mathbb{E}} [f(\theta)] .$$

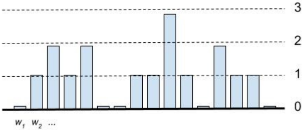
Original weights:

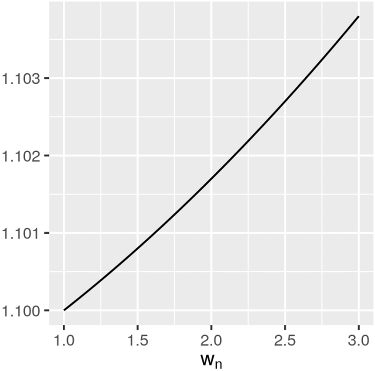


Leave-one-out weights:



Bootstrap weights:



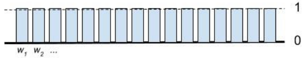


Data re-weighting.

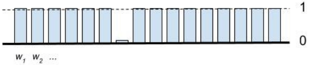
Augment the problem with *data weights* $w_1, \dots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \quad \Rightarrow \text{We write} \underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)].$$

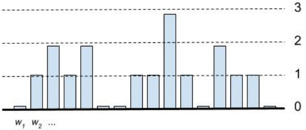
Original weights:

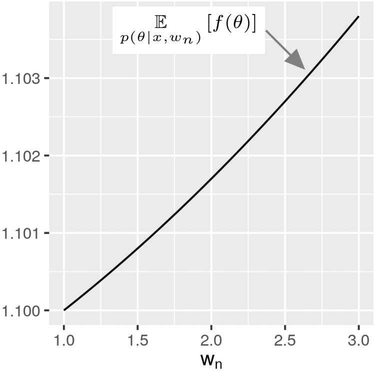


Leave-one-out weights:



Bootstrap weights:



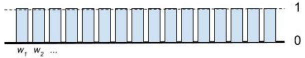


Data re-weighting.

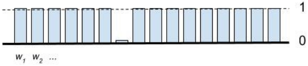
Augment the problem with *data weights* $w_1, \ldots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)].$$

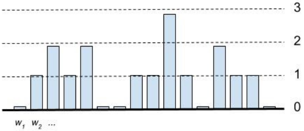
Original weights:

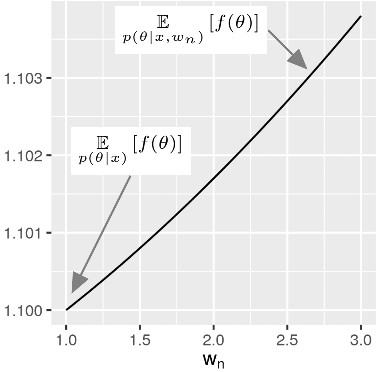


Leave-one-out weights:



Bootstrap weights:



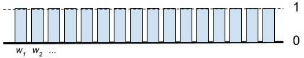


Data re-weighting.

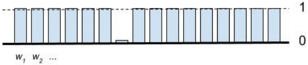
Augment the problem with *data weights* $w_1, \ldots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \quad \Rightarrow \text{We write} \underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)].$$

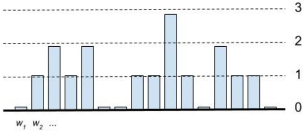
Original weights:

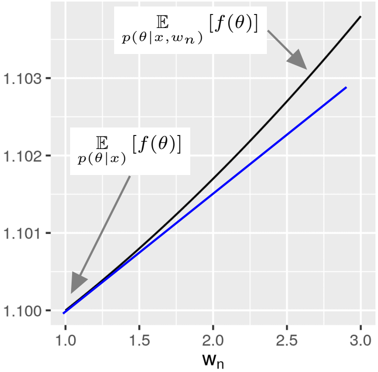


Leave-one-out weights:



Bootstrap weights:



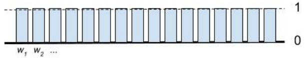


Data re-weighting.

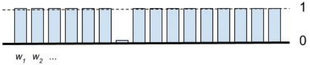
Augment the problem with *data weights* $w_1, \ldots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \quad \Rightarrow \text{We write} \underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)].$$

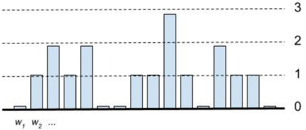
Original weights:

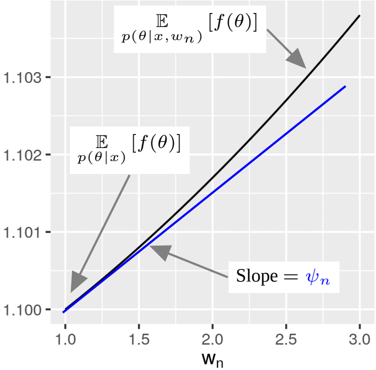


Leave-one-out weights:



Bootstrap weights:



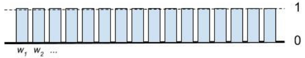


Data re-weighting.

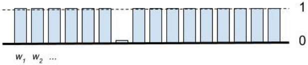
Augment the problem with *data weights* $w_1, \ldots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \qquad \Rightarrow \text{We write } \mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)].$$

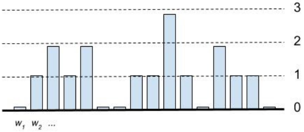
Original weights:

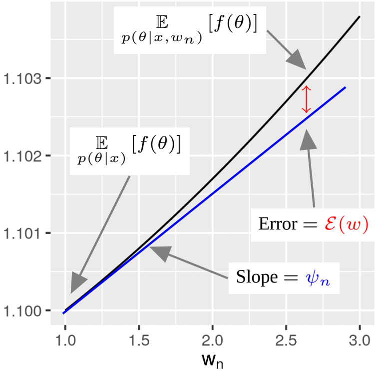


Leave-one-out weights:



Bootstrap weights:



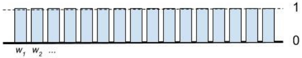


Data re-weighting.

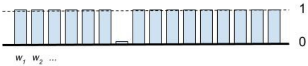
Augment the problem with *data weights* $w_1, \ldots, w_N$.

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^{N} w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)].$$

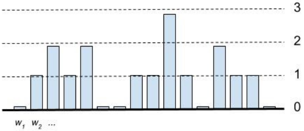
Original weights:

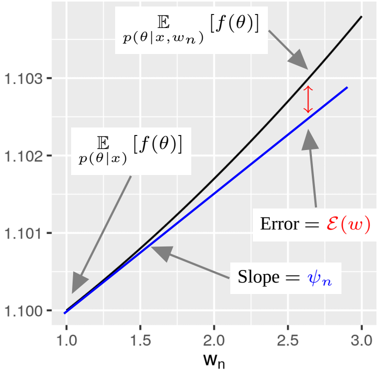


Leave-one-out weights:



Bootstrap weights:





The re-scaled slope $N\psi_n$ is known as the “influence function” at data point x_n.

$$\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w).$$

Data re-weighting.

How can we use the approximation?

Example: Approximate bootstrap.

Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

This is equivalent to re–sampling data with replacement.

Data re-weighting.

How can we use the approximation?

Example: Approximate bootstrap.

Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

This is equivalent to re–sampling data with replacement.

$$\begin{aligned}
\text{Bootstrap variance} &= \underset{p(w)}{\text{Var}}\left(\underset{p(\theta|x,w)}{\mathbb{E}}[f(\theta)]\right) \\
&\approx \underset{p(w)}{\text{Var}}\left(\underset{p(\theta|x)}{\mathbb{E}}[f(\theta)] + \psi_n(w_n - 1)\right) \qquad \text{(assuming the error is small)} \\
&= \sum_{n=1}^{N}\left(\psi_n - \overline{\psi}\right)^2.
\end{aligned}$$

The final line is also known as the "infinitesimal jackknife" variance approximation.

Data re-weighting.

How can we use the approximation?

Example: Approximate bootstrap.

$$\text{Bootstrap variance} \approx \sum_{n=1}^{N} \left(\psi_n - \overline{\psi} \right)^2 .$$

Data re-weighting.

How can we use the approximation?

Example: Approximate bootstrap.

$$\text{Bootstrap variance} \approx \sum_{n=1}^{N} \left(\psi_n - \overline{\psi}\right)^2.$$

Example: Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathop{\mathbb{E}}_{p(\theta|x,w_{(-n)})} [-\ell(x_n|\theta)] \approx \mathop{\mathbb{E}}_{p(\theta|x)} [-\ell(x_n|\theta)] - \psi_n$$

Data re-weighting.

How can we use the approximation?

Example: Approximate bootstrap.

$$\text{Bootstrap variance} \approx \sum_{n=1}^{N} \left(\psi_n - \overline{\psi} \right)^2 .$$

Example: Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathop{\mathbb{E}}_{p(\theta|x,w_{(-n)})} \left[-\ell(x_n|\theta) \right] \approx \mathop{\mathbb{E}}_{p(\theta|x)} \left[-\ell(x_n|\theta) \right] - \psi_n$$

Example: Influential subsets: Approximate maximum influence perturbation (AMIP).

Let $W_{(-K)}$ denote weights leaving out K points.

$$\max_{w \in W_{(-K)}} \left(\mathop{\mathbb{E}}_{p(\theta|x,w)} \left[f(\theta) \right] - \mathop{\mathbb{E}}_{p(\theta|x)} \left[f(\theta) \right] \right) \approx - \sum_{n=1}^{K} \psi_{(n)} .$$

Data re-weighting.

How can we use the approximation?

Example: Approximate bootstrap.

$$\text{Bootstrap variance} \approx \sum_{n=1}^{N} \left(\psi_n - \overline{\psi} \right)^2 .$$

Example: Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathop{\mathbb{E}}_{p(\theta|x,w_{(-n)})} \left[-\ell(x_n|\theta) \right] \approx \mathop{\mathbb{E}}_{p(\theta|x)} \left[-\ell(x_n|\theta) \right] - \psi_n$$

Example: Influential subsets: Approximate maximum influence perturbation (AMIP).

Let $W_{(-K)}$ denote weights leaving out K points.

$$\max_{w \in W_{(-K)}} \left(\mathop{\mathbb{E}}_{p(\theta|x,w)} \left[f(\theta) \right] - \mathop{\mathbb{E}}_{p(\theta|x)} \left[f(\theta) \right] \right) \approx -\sum_{n=1}^{K} \psi_{(n)} .$$

How to compute the slopes ψ_n?
How large is the error $\mathcal{E}(w)$?

Expressions for the slope and error

How to compute the slopes ψ_n? How can we analyze the error $\mathcal{E}(w)$?

$$\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w).$$

Expressions for the slope and error

How to compute the slopes ψ_n? How can we analyze the error $\mathcal{E}(w)$?

$$\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w).$$

By dominated convergence, $\psi_n = \underbrace{\underset{p(\theta|X)}{\mathrm{Cov}}(f(\theta), \ell_n(\theta))}_{\text{Estimatable with MCMC!}}$.

Expressions for the slope and error

How to compute the slopes ψ_n? How can we analyze the error $\mathcal{E}(w)$?

$$\underset{p(\theta|X,w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}(w).$$

By dominated convergence, $\psi_n = \underbrace{\underset{p(\theta|X)}{\mathrm{Cov}}(f(\theta), \ell_n(\theta))}_{\text{Estimatable with MCMC!}}.$

Furthermore, by the mean value theorem, for some $\tilde{w}$,

$$\mathcal{E}(w) = \frac{1}{2} \sum_{n=1}^{N} \sum_{n'=1}^{N} \mathcal{E}_{nn'}(w)(w_n - 1)(w_{n'} - 1) \quad \text{where}$$

$$\mathcal{E}_{nn'}(w) := \underbrace{\underset{p(\theta|X,\tilde{w})}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\bar{\ell}_{n'}(\theta)\right]}_{\substack{\text{Cannot compute directly!} \\ \text{(we don't know the intermediate value theorem's } \tilde{w}\text{).}}}$$

But we can analyze it.

Here, an overbar denotes "posterior–mean zero." For example, $\bar{f}(\theta) := f(\theta) - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Theoretical results

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of $\sqrt{N} \underset{p(\theta|X)}{\mathbb{E}} [f(\theta)]$?

Theoretical results

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of $\sqrt{N} \underset{p(\theta|X)}{\mathbb{E}} [f(\theta)]$?

Theorem 3 of Giordano and Broderick [2023] (paraphrase):
If the parameter dimension is fixed, and Berstein–von Mises (BVM) theorem–like conditions hold, then the IJ covariance is consistent, because $\sqrt{N}\mathcal{E}(w) \xrightarrow[N\to\infty]{prob} 0$.

Theoretical results

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of $\sqrt{N} \underset{p(\theta|X)}{\mathbb{E}} [f(\theta)]$?

Theorem 3 of Giordano and Broderick [2023] (paraphrase):
If the parameter dimension is fixed, and Berstein–von Mises (BVM) theorem–like conditions hold, then the IJ covariance is consistent, because $\sqrt{N}\mathcal{E}(w) \xrightarrow[N\to\infty]{prob} 0$.

Problem: we're doing MCMC because BVM does not hold.
What if $f(\theta)$ concentrates marginally, but some components don't concentrate?

Theoretical results

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of $\sqrt{N} \underset{p(\theta|X)}{\mathbb{E}} [f(\theta)]$?

Theorem 3 of Giordano and Broderick [2023] (paraphrase):
If the parameter dimension is fixed, and Berstein–von Mises (BVM) theorem–like conditions hold, then the IJ covariance is consistent, because $\sqrt{N}\mathcal{E}(w) \xrightarrow[N\to\infty]{prob} 0$.

Problem: we're doing MCMC because BVM does not hold.
What if $f(\theta)$ concentrates marginally, but some components don't concentrate?

Theorem 4 of Giordano and Broderick [2023] (paraphrase & conjecture):
In a flexible class of high–dimensional exponential family models,
even when $p\left(f(\theta)|X\right)$ obeys a BVM marginally (!),

- $\sqrt{N}\mathcal{E}(w)$ does not converge to zero (so the IJ covariance is inconsistent), but...
- $\sqrt{N}\mathcal{E}(w) = \tilde{O}_p\left(1\right)$, and proportional to the nuisance parameters' posterior covariance

Theoretical results

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of $\sqrt{N} \underset{p(\theta|X)}{\mathbb{E}} [f(\theta)]$?

Theorem 3 of Giordano and Broderick [2023] (paraphrase):
If the parameter dimension is fixed, and Berstein–von Mises (BVM) theorem–like conditions hold, then the IJ covariance is consistent, because $\sqrt{N}\mathcal{E}(w) \xrightarrow[N\to\infty]{prob} 0$.

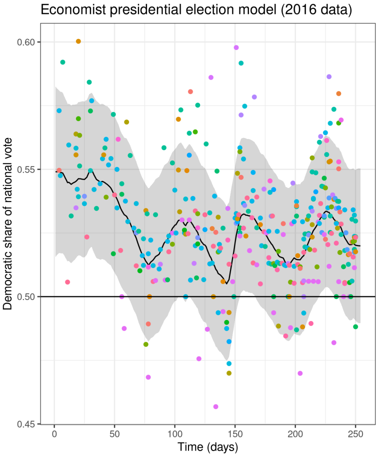
Problem: we're doing MCMC because BVM does not hold.
What if $f(\theta)$ concentrates marginally, but some components don't concentrate?

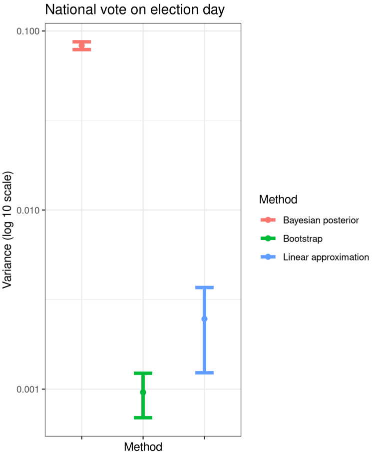
Theorem 4 of Giordano and Broderick [2023] (paraphrase & conjecture):
In a flexible class of high–dimensional exponential family models,
even when $p\left(f(\theta)|X\right)$ **obeys a BVM marginally (!),**

- $\sqrt{N}\mathcal{E}(w)$ does not converge to zero (so the IJ covariance is inconsistent), but...
- $\sqrt{N}\mathcal{E}(w) = \tilde{O}_p\left(1\right)$, and proportional to the nuisance parameters' posterior covariance

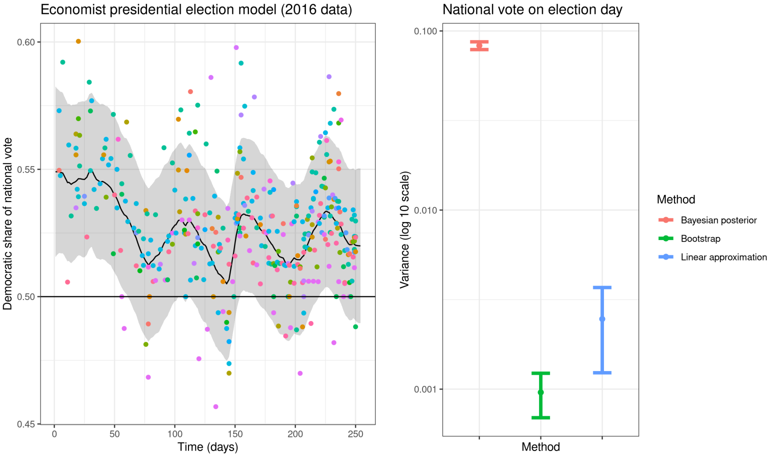
$\Rightarrow$ **High–dimensional Bayesian models are an extremely common class of problems for which the influence function may not provide a good approximation.**

Observations and consequences





Observations and consequences



Preprint: Giordano and Broderick [2023] (`arXiv:2305.06466`)

- Detailed proofs
- Simple analytical examples
- Simulated and real–world experiments

References

A. Gelman and M. Heidemanns. The Economist: Forecasting the US elections., 2020. URL `https://projects.economist.com/us-2020-forecast/president`. Data and model accessed Oct., 2020.

R. Giordano and T. Broderick. The Bayesian infinitesimal jackknife for variance. *arXiv preprint arXiv:2305.06466*, 2023.

J. Huggins and J. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79–104, 2023.

Data re-weighting.

How can we use the approximation?

Data re-weighting.

How can we use the approximation?

Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \underset{p(\theta|x, w_{(-n)})}{\mathbb{E}}[f(\theta)] \approx \underset{p(\theta|x)}{\mathbb{E}}[f(\theta)] - \psi_n$$

Data re-weighting.

How can we use the approximation?

Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathop{\mathbb{E}}_{p(\theta|x,w_{(-n)})}[f(\theta)] \approx \mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)] - \psi_n$$

Example: Approximate bootstrap.

Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\begin{aligned}\text{Bootstrap variance} &= \mathop{\text{Var}}_{p(w)}\left(\mathop{\mathbb{E}}_{p(\theta|x,w)}[f(\theta)]\right) \\ &\approx \mathop{\text{Var}}_{p(w)}\left(\mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)] + \psi_n(w_n - 1)\right) \\ &= \sum_{n=1}^{N}\left(\psi_n - \overline{\psi}\right)^2.\end{aligned}$$

Data re-weighting.

How can we use the approximation?

Cross validation. Let $w_{(-n)}$ leave out point n, and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathop{\mathbb{E}}_{p(\theta|x,w_{(-n)})}[f(\theta)] \approx \mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)] - \psi_n$$

Example: Approximate bootstrap.

Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\begin{aligned}
\text{Bootstrap variance} &= \mathop{\text{Var}}_{p(w)}\left(\mathop{\mathbb{E}}_{p(\theta|x,w)}[f(\theta)]\right) \\
&\approx \mathop{\text{Var}}_{p(w)}\left(\mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)] + \psi_n(w_n - 1)\right) \\
&= \sum_{n=1}^{N}\left(\psi_n - \overline{\psi}\right)^2.
\end{aligned}$$

Influential subsets: Approximate maximum influence perturbation (AMIP).

Let $W_{(-K)}$ denote weights leaving out K points.

$$\max_{w \in W_{(-K)}}\left(\mathop{\mathbb{E}}_{p(\theta|x,w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|x)}[f(\theta)]\right) \approx -\sum_{n=1}^{K}\psi_{(n)}.$$

Example: A negative binomial model

Consider $p(X|\gamma) = \prod_{n=1}^{N} \text{NegativeBinomial}(x_n|\gamma)$. Here, $\theta = \gamma$ is a scalar.

Example: A negative binomial model

Consider $p(X|\gamma) = \prod_{n=1}^{N} \text{NegativeBinomial}(x_n|\gamma)$. Here, $\theta = \gamma$ is a scalar.

As $N \to \infty$, $p(\gamma|X)$ concentrates at rate $1/\sqrt{N}$ (Bernstein–von Mises).

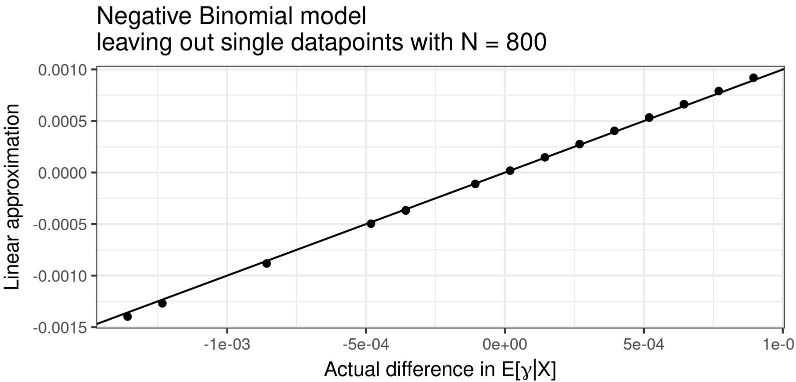
$$\Rightarrow N \left(\underset{p(\gamma|X, w_n)}{\mathbb{E}} [\gamma] - \underset{p(\gamma|X)}{\mathbb{E}} [\gamma] \right) = \psi_n (w_n - 1) + O_p(N^{-1}).$$

Example: A negative binomial model

Consider $p(X|\gamma) = \prod_{n=1}^{N} \mathrm{NegativeBinomial}(x_n|\gamma)$. Here, $\theta = \gamma$ is a scalar.

As $N \to \infty$, $p(\gamma|X)$ concentrates at rate $1/\sqrt{N}$ (Bernstein–von Mises).

$$\Rightarrow N \left(\underset{p(\gamma|X,w_n)}{\mathbb{E}} [\gamma] - \underset{p(\gamma|X)}{\mathbb{E}} [\gamma] \right) = \color{blue}{\psi_n}(w_n - 1) + \color{red}{O_p(N^{-1})}.$$

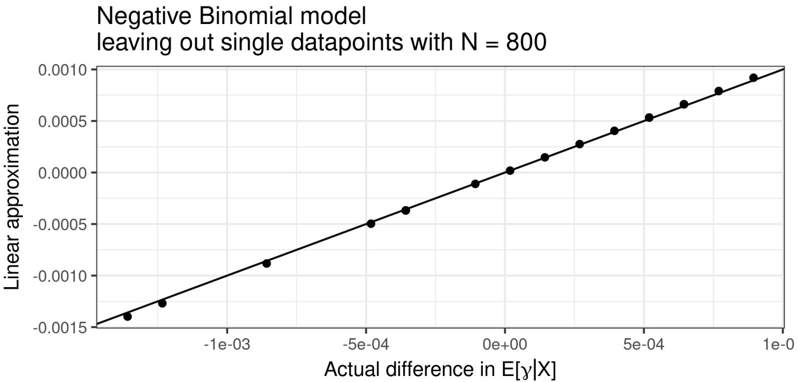


Example: A negative binomial model

Consider $p(X|\gamma) = \prod_{n=1}^{N} \mathrm{NegativeBinomial}(x_n|\gamma)$. Here, $\theta = \gamma$ is a scalar.

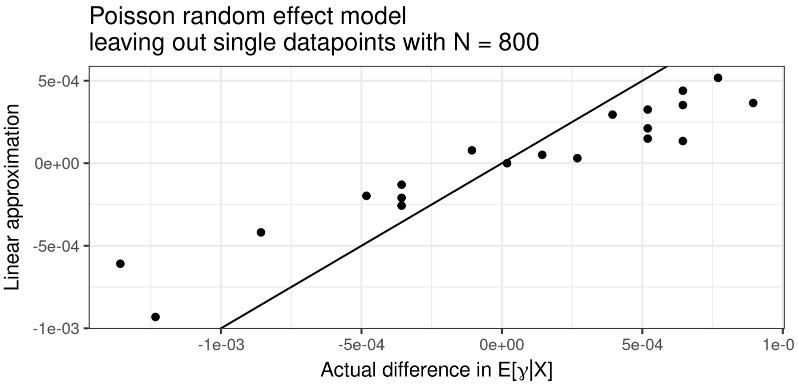
As $N \to \infty$, $p(\gamma|X)$ concentrates at rate $1/\sqrt{N}$ (Bernstein–von Mises).

$$\Rightarrow N \left(\underset{p(\gamma|X, w_n)}{\mathbb{E}} [\gamma] - \underset{p(\gamma|X)}{\mathbb{E}} [\gamma] \right) = \psi_n (w_n - 1) + O_p(N^{-1}).$$



Problem: Most computationally hard Bayesian problems don't concentrate.

Example: **Poisson model with random effects (REs) λ and fixed effect γ.**



A contradiction?

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

A contradiction?

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

With a constant regressor, Gamma REs, and one RE per observation, these are the same model, with the same $p(\gamma|X)$.

Is $\underset{p(\gamma|X,w)}{\mathbb{E}}[\gamma]$ linear in the data weights or not?

A contradiction?

Negative binomial observations.

Asymptotically linear in w.

$$\log p(X|\gamma, w^m) = \sum_{n=1}^{N} w_n^m \log p(x_n|\gamma)$$

Poisson observations with random effects.

Asymptotically non-linear in w.

$$\log p(X|\gamma, \lambda, w^c) = \sum_{n=1}^{N} w_n^c \log p(x_n|\lambda, \gamma)$$

With a constant regressor, Gamma REs, and one RE per observation, these are the same model, with the same $p(\gamma|X)$.

Is $\underset{p(\gamma|X,w)}{\mathbb{E}}[\gamma]$ linear in the data weights or not?

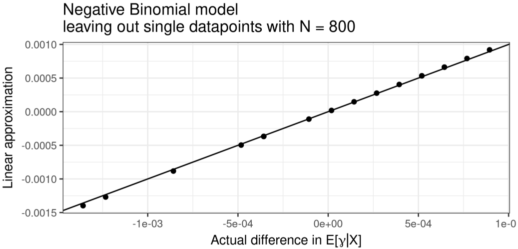
Trick question! We weight a log likelihood contribution, not a datapoint.

The two weightings are not equivalent in general.

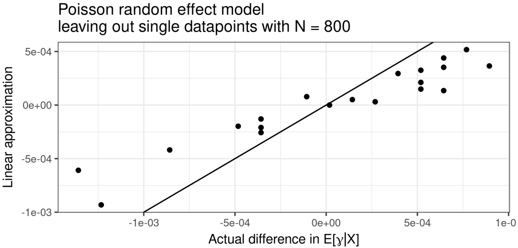
Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Uses $\log p(x_n|\gamma)$:
$$\psi_n = \mathop{\mathbb{E}}_{p(\gamma|X)} \left[\bar{\gamma}\bar{\ell}_n(\gamma)\right]$$



Uses $\log p(x_n|\gamma, \lambda)$:
$$\psi_n = \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)} \left[\bar{\gamma}\bar{\ell}_n(\gamma, \lambda)\right]$$

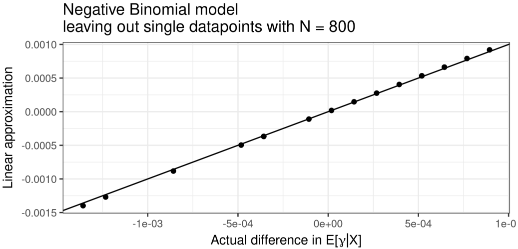


Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

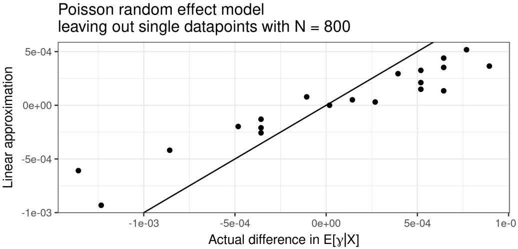
Uses $\log p(x_n|\gamma)$:
$$\psi_n = \mathop{\mathbb{E}}_{p(\gamma|X)} \left[\bar{\gamma}\bar{\ell}_n(\gamma)\right]$$

Not computable from
$\gamma, \lambda \sim p(\gamma, \lambda | X)$
in general.



Uses $\log p(x_n|\gamma, \lambda)$:
$$\psi_n = \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)} \left[\bar{\gamma}\bar{\ell}_n(\gamma, \lambda)\right]$$

Computable from
$\gamma, \lambda \sim p(\gamma, \lambda | X)$.

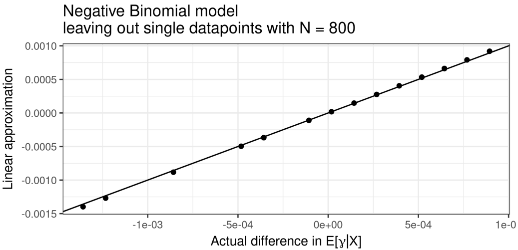


Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Uses $\log p(x_n|\gamma)$:
$$\psi_n = \underset{p(\gamma|X)}{\mathbb{E}} \left[\bar{\gamma}\bar{\ell}_n(\gamma)\right]$$

Not computable from
$\gamma, \lambda \sim p(\gamma, \lambda|X)$
in general.



Uses $\log p(x_n|\gamma, \lambda)$:
$$\psi_n = \underset{p(\gamma,\lambda|X)}{\mathbb{E}} \left[\bar{\gamma}\bar{\ell}_n(\gamma, \lambda)\right]$$

Computable from
$\gamma, \lambda \sim p(\gamma, \lambda|X)$.

May still be useful when $p(\lambda|X)$ is *somewhat* concentrated.